Evaluating Creativity in Large Language Models through Creative Problem-Solving: A New Dataset and Benchmark

Anonymous ACL submission

Abstract

Creative problem-solving, integrating divergent and convergent thinking, is pivotal for leveraging creativity in fields such as AI4Science. As large language models (LLMs) evolve into sophisticated creative assistants, it becomes crucial to effectively assess their problem-solving abilities. Traditional benchmarks, often rooted in cognitive science, focus on a single phase or do not distinguish between the divergent and convergent phases, limiting their ability to fully evaluate LLMs. To bridge this gap, we introduce a novel benchmark comprising an open-ended question answering (QA) dataset alongside traditional creativity tasks, aimed at evaluating the holistic creative capabilities of LLMs. This benchmark utilizes multi-dimensional evaluation metrics to provide a comprehensive assessment that correlates with model parameters, architectural differences, and domain-specific expertise. The benchmark aims to not only advance understanding in the field but also set a new standard for evaluating the creative problem-solving potential of LLMs. The dataset and code are available at: https://anonymous.4open.science/r/LLM-creativity-Benchmark/.

1 Introduction

Creativity, a pivotal research topic within cognitive science, plays an essential role in enhancing our understanding of human behavior, cognitive processes, and innovation capacity across various domains including arts and sciences. The exploration of creativity extends beyond mere theoretical inquiry, influencing practical applications and technological advancements.

Within the broad spectrum of creativity research, creative problem-solving (CPS) emerges as a critical focus. This field particularly emphasizes the synthesis of divergent thinking—generating a multitude of potential solutions—and convergent thinking—implementing the most effective solutions (Couger et al., 1993; Guilford, 2017). CPS is thus integral to developing processes that enhance innovation by effectively combining expansive ideation with focused problem resolution.

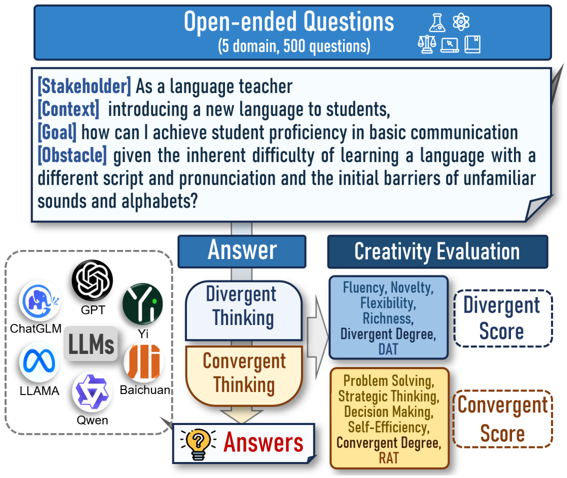

Figure 1: The overall framework of the creative-problem solving benchmark.

In the context of large language models (LLMs), the relevance of CPS skills has notably increased, marking a pivotal advancement towards Artificial General Intelligence (AGI). Models such as GPT-4 (Achiam et al., 2023) have showcased profound capabilities in generating complex, contextually relevant content, thereby serving as creative collaborators. This is particularly evident in sectors like consulting and AI-driven scientific research, for instance, AI4Science, where they facilitate innovative problem-solving and decision-making processes (Anishchenko et al., 2021).

Recent studies have ventured into exploring and enhancing the CPS capabilities of LLMs, scrutinizing their performance across a diverse array of tasks. These tasks range from the Alternate Uses Task (AUT) (Tian et al., 2023), humor analysis (Zhong et al., 2023), and Divergent Thinking Assessment (DAT) (Olson et al., 2021), to the Remote Associates Test (RAT) (Mednick, 1962), Torrance Tests of Creative Thinking (TTCT) (Guzik et al., 2023), and specialized applications like protein design (Anishchenko et al., 2021). Each of these tasks contributes to a broader understanding of the creative spectrum these models can engage, offering insights into their versatility and adaptability in generating innovative solutions.

Despite the progress, evaluating the CPS proficiency of LLMs remains fraught with challenges. The traditional assessment methods such as DAT, AUT, TTCT, and RAT, primarily designed for human evaluation, often fail to capture the unique problem-solving dynamics inherent in LLMs. Moreover, the inherent complexity of CPS, which requires the integration of divergent and convergent thinking processes, is not fully addressed as most studies focus predominantly on one phase over the other. This oversight restricts a holistic assessment of LLMs' capabilities in creative tasks. Additionally, the scarcity of open-source datasets and lack of standardized metrics further complicate the evaluation landscape, significantly impacting the ability to measure and optimize LLMs across various model architectures and settings.

To bridge these gaps, our study embarks on a comprehensive evaluation of CPS within LLMs. It begins with an extensive review of existing research on LLM-based creative methodologies, clarifying the definitions and expanding the scope of CPS within this specific context, aiming to set the stage for a deeper understanding of how creative processes can be measured and enhanced in LLMs. Following this, a novel open-source dataset focused on open-ended questions, encompassing both general and domain-specific inquiries, has been developed. This dataset is meticulously crafted to rigorously evaluate the creative capabilities of LLMs, fostering a more nuanced understanding of their potential.

Building on this dataset, a benchmark has been constructed to assess CPS in LLMs, integrating a set of multi-view assessment metrics. These metrics are tailored to evaluate both traditional creative tasks and open-ended question responses, facilitating a comprehensive examination of LLM performance. Extensive experimentation using this benchmark has allowed for a thorough evaluation of various LLMs, elucidating their strengths and weaknesses in handling CPS tasks.

Finally, drawing on empirical findings and theoretical insights, strategies aimed at refining the CPS capacities of LLMs have been formulated. These strategies are designed to enhance both the divergent and convergent thinking abilities inherent in LLMs, striving for a more balanced and effective output in creative problem-solving. Subsequent experiments have confirmed the efficacy of these optimization strategies, showcasing noticeable improvements in the performance of LLMs across a spectrum of CPS tasks.

2 Preliminary and Related Work

2.1 Definition of Creativity

The concept of creativity has been defined in myriad ways, with over 100 different definitions identified in the literature (Treffinger, 1998). However, the vast majority of studies on creativity tend to focus on a small subset of these definitions. In cognitive science, creativity is commonly examined from four distinct angles: the cognitive processes involved in creative thinking (referred to as 'process' in this paper), the traits of creative individuals ('person'), the outcomes of creative efforts ('product'), and the interplay between a creative individual and their environment ('press') (Couger et al., 1993). This paper concentrates on the 'process' and 'product' aspects of creativity as they are most relevant to the analysis of LLMs, while the 'person' and 'press' aspects are more pertinent to studies of human creativity.

The process perspective of creativity, as defined by (Torrance, 1977), involves recognizing problems or knowledge gaps, formulating hypotheses, testing and validating these hypotheses, and sharing the results. Another perspective by (Mednick, 1962) suggests that creativity entails merging associative elements into new configurations that meet the demands of a specific task. Additionally, (Guilford, 2017) describes creativity as a problem-solving activity, distinguishing between divergent and convergent cognitive operations. Divergent production is marked by a broad search for various logical solutions to open-ended issues, whereas convergent production focuses on a narrow search for a single, precise answer to a specific problem, highlighting that divergent processes are more closely linked to effective creative thinking.

From the product-oriented perspective, (Khatena and Torrance, 1973) views creativity as the ability to construct or organize ideas, thoughts, and emotions into unusual and associative links through imaginative power. (Gardner, 2011) argues that creative individuals are capable of solving problems, creating products, or posing new questions in ways that are both novel and culturally appropriate. Creativity is also seen as the capability to generate or devise something original and suitable to task constraints, which is also high in quality, useful, aesthetically appealing, and novel.

Despite the diversity of perspectives, this paper follows the definition by (Guilford, 2017), and proposes a formal definition of creativity tailored to the task characteristics of LLMs. Based on the prevailing definitions in cognitive science and creativity studies, we define creativity as the capacity to generate diverse and novel ideas or solutions during the divergence phase, followed by the ability to refine and select the most valuable and applicable ones during the convergence phase. Specifically, *divergent thinking* refers to the process of generating a wide array of possible ideas, solutions, or associations without immediate constraints on feasibility or practicality. This is particularly crucial for the initial phase of creative tasks where potential is maximized. On the other hand, *convergent thinking* involves the critical evaluation and narrowing down of choices to identify the most effective, practical, and innovative outcomes. This two-phase approach allows for a comprehensive assessment of an LLM's creativity, capturing both its generative and evaluative capacities. Thus, creativity in LLMs can be conceptualized as the interplay and balance between these two cognitive phases, enabling the generation of solutions that are not only original but also appropriate and useful for given constraints.

2.2 Related Work

2.2.1 Approaches for Measuring Creativity

Measuring creativity within the domain of cognitive science presents considerable challenges, primarily due to its subjective nature and the diverse environments in which it manifests. Among the myriad approaches developed to quantify creativity, this section focuses on the process and product dimensions, particularly highlighting the Torrance Tests of Creative Thinking (TTCT) (Torrance, 1977), Divergent Thinking Tests (DAT) (Olson et al., 2021), and the Remote Associates Test (RAT) (Mednick, 1962). The TTCT and DAT are instrumental in assessing the creative process. These tests measure ideational fluency through tasks that require participants to generate as many responses as possible to open-ended questions. The responses are evaluated based on fluency (the number of responses), originality (statistical rarity of the responses), flexibility (variety of categories the responses fall into), and elaboration (detail of the responses). Such divergent thinking tests are designed not just to gauge the quantity but also the quality of creative responses, reflecting an individual's capacity to navigate through ill-structured problems creatively. On the other hand, the RAT focuses on convergent thinking by evaluating the ability to form novel and useful combinations from seemingly unrelated elements. This task challenges participants to bridge associative gaps, reflecting a different dimension of creative thought that emphasizes synthesis over generation.

Despite their widespread use, the psychometric foundations and cognitive underpinnings of these tests, particularly convergent thinking tasks, continue to stir debate within the research community.

2.2.2 Research on Creativity in Large Language Models

The aforementioned section outlines methodologies for measuring human creativity within the domain of cognitive science. However, due to inherent differences between LLMs and humans, these traditional methods may lead to irrelevant or logically flawed responses when applied to LLMs. This discrepancy necessitates a critical examination of these methods, leading us to question: *How can we adapt these measures to effectively evaluate the creativity of LLMs?* Given the burgeoning potential of LLMs, researchers have explored two primary types of approaches for assessing their creativity.

The first approach involves adapting established cognitive science techniques to LLM contexts. For instance, (Stevenson et al., 2022) utilized the Alternative Uses Task (AUT) to compare the creative outputs of GPT-3 with those of humans, finding that humans generally produced more creative responses. Further, (Summers-Stay et al., 2023) refined this approach by evaluating the originality and practicality of responses previously generated by GPT-3. Despite GPT-3's ability to generate compelling ideas, it struggled with discarding impractical ones. Another study by (Naeini et al., 2023) curated a dataset from the British quiz show *Only Connect*, serving as an analogical proxy for RAT tasks, to assess creative problem-solving in LLMs. (Cropley, 2023; Chen and Ding, 2023) compared the creativity of GPT-4 and GPT-3.5 using DAT against human norms. Contrarily, (Góes et al.,

2023) introduced an interactive method that enables GPT-4 to autonomously refine its creative outputs, employing both AUT and TTCT visual completion tasks.

The second approach involves devising entirely new methodologies tailored for LLMs. (Wang et al., 2024) theoretically demonstrated that LLMs could achieve human-level creativity by fitting data generated by human creators. They introduced concepts of 'relative creativity'—where an LLM is considered as creative as a realistic human creator if its outputs are indistinguishable by an evaluator—and 'statistical creativity', which assesses how an LLM's creativity compares to existing human creators. Furthermore, (Lee, 2023) developed a mathematical framework to explore the trade-off between hallucination and creativity in LLMs, providing a rigorous analysis of the phenomenon.

Despite the growing interest in LLM creativity, the field requires more comprehensive benchmarks to deepen our understanding and enhance the assessment of creativity in LLMs.

3 Dataset Construction

For the Open-ended QA dataset, we have adopted the domain selections from the previous research (Li et al., 2024), incorporating a general domain to cover a wide array of topics alongside four representative specialized domains: Finance, Science, Education, and Biology. We employ GPT-4 as an examiner to generate diverse and high-quality questions across these domains. For each domain, GPT-4 is prompted to produce 100 unique questions. However, the varying capabilities of GPT-4 across different specialized domains raise important considerations regarding the consistency of question quality. These discrepancies are likely due to the model's inherent strengths and weaknesses in handling domain-specific knowledge, which can significantly impact the quality and relevance of the questions it generates.

Inspired by previous research detailed in (Ding et al., 2023), we have designed a structured prompt approach that divides each question into four components: the stakeholder (the entity the question is directed towards or about), the context (the scenario or background information relevant to the question), the goal (what the question aims to achieve or uncover), and the obstacle (any challenges or complications inherent to the question). This structured prompt approach is designed to foster clearer, more targeted, and ultimately higher-quality questions by aligning them more closely with real-world issues and theoretical considerations. Furthermore, our prompt incorporates few-shot learning, a technique that involves presenting the model with a few examples within the prompt, thereby enhancing the quality of the questions it generates.

Additionally, we modify the prompt every 20 questions during the question generation process. Specifically, since we structure the questions into four components and utilize few-shot learning, we alter the prompt to either closely align with or greatly differ from the components in the few-shot examples. This approach helps to ensure that the questions generated are as diverse as possible. Finally, we employ GPT-4 to reorganize and rewrite the four components into a cohesive and logically structured question, the examples are shown in Table 1 The detailed prompt examples can be found in Appendix A.1.

4 LLM Creativity Benchmark

In this section, we discuss the methodology for evaluating LLM creativity, including the tasks of Open-ended Question Answering (Open-ended QA), DAT, and RAT. This includes the construction of specifically chosen datasets and the design of evaluation metrics to assess divergent and convergent thinking capabilities of creativity.

4.1 Experiment Settings

4.1.1 Evaluation Tasks and Datasets

The benchmark is structured to evaluate the creativity of LLMs across divergent and convergent thinking stages using three tasks: Open-ended QA, DAT, and RAT. Each task is chosen for its relevance to specific cognitive processes and is supported by carefully selected datasets.

The Open-ended QA task is central to our evaluation framework as it simultaneously assesses LLMs' abilities in both divergent and convergent thinking. This task requires models to generate innovative solutions to novel problems, thereby testing their creative output in unstructured scenarios. The dataset for Open-ended QA is a multi-domain collection that we have specifically curated in Section 3. It presents a broad range of challenges designed to elicit original and practical solutions from the models, ensuring a comprehensive test of their ability to navigate and respond to diverse queries.

Prompt	Please generate a new situation in **{Finance, Science, Education, Biology, General}** domain that has a very different **(stakeholder, context)** but very similar **(goal, obstacle)** based on the input:.
Few-shot example	**Stakeholder**: a patient who has a malignant tumor in his stomach **Context**: ray at low intensity is insufficient to destroy the tumor **Goal**: destroy the tumor without affecting the healthy tissue **Obstacle**: ray at high intensity will also destroy healthy tissue
Finance	As a financial analyst at a renewable energy company dealing with fluctuating energy market prices, how can I maximize investment returns on renewable projects given the unpredictability of government policies affecting energy sector profitability?
Education	As a math tutor working with students who have math anxiety, how can I help students gain confidence and improve their math skills while making math accessible and engaging to alleviate anxiety and build competence?
Science	As a pharmacologist developing new antiviral drugs, how can I create effective treatments for emerging viral infections while balancing drug efficacy with minimal side effects in a rapidly changing viral landscape?
Biology	As an allergist developing a new allergy immunotherapy, how can I provide long-term relief from allergies without the treatment triggering severe allergic reactions during the desensitization process?

Table 1: Generated example of different domain questions

Together, these tasks and their corresponding datasets provide a comprehensive framework for assessing the creative capacities of LLMs across different dimensions of thought.

4.1.2 Model Selection

This study encompasses a diverse range of LLMs to ensure a comprehensive evaluation of their creative capabilities. The selection of models is based on several criteria, including model architecture and parameter count. In the closed-source category, we have chosen widely used models such as GPT-3.5 and GPT-4, which represent some of the most advanced capabilities in LLM technology. Their inclusion is crucial for benchmarking state-of-the-art performance in creativity tasks within proprietary models. For open-source models, our selection is guided by the popularity and usage metrics from repositories like Hugging Face, ensuring that the models included, such as LLAMA-2 and Yi, are not only representative of current community engagement but also of varied architectural approaches. Specifically, we have included multiple versions of LLAMA-2 (i.e., 7b, 13b, and 70b) and Yi (i.e., 6b and 34b) to analyze the impact of model size on creative output. Additionally, models like Qwen1.5-14b, BaiChuan2-13b, and Chatglm2-6b are chosen to broaden the evaluation spectrum further, allowing us to explore how different training methodologies and design principles affect creative performance. This varied selection of models, spanning different architectures and sizes, provides a robust foundation for assessing and comparing the creative capabilities of LLMs under a standardized set of tasks and metrics.

4.1.3 Evaluation Metrics

Several methods are commonly used to evaluate QA tasks within LLMs, notably including Likert scale scoring (Joshi et al., 2015). In developing our benchmark, we are inspired by the Likert scale method and the established framework from creativity research in cognitive science, as discussed in (Boden, 1994). We have devised a set of metrics specifically designed to evaluate the creativity of LLMs. Our creativity metrics function as an absolute evaluative measure, where the evaluator assigns scores to a given response along predefined dimensions. We have identified two main aspects of creativity and established four distinct dimensions within each aspect of our dataset.

For individual answer evaluation, we assess the divergent and convergent thinking abilities of LLMs through carefully chosen metrics. For divergent thinking, we measure *Fluency*, *Novelty*, *Flexibility*, and *Richness*. Each of these metrics serves a specific purpose: *Fluency* quantifies the volume of ideas, *Novelty* evaluates the uniqueness, *Flexibility* assesses the variety across categories, and *Richness* gauges the depth of the ideas, as supported by studies such as (Guzik et al., 2023; Zhao et al., 2024). For convergent thinking, we apply metrics including *Problem Solving*, *Strategic Thinking*, *Decision Making*, and *Self-Efficiency*, which are chosen based on their emphasis in recent cognitive research (Du, 2023), ensuring that each metric contributes to a comprehensive understanding of how LLMs manage and optimize creative outputs. All of these metrics are scored on a scale of 1 to 10, ranging from worst to best.

Moving beyond single-answer analysis, we compute a *Divergence Degree* and a *Convergence Degree* for each model from multiple responses, aiming to not only evaluate isolated instances of creativity but also to understand the broader creative process. The final scores for each dimension of the model are calculated as the average of all problem scores. Both of the two metrics are scored on a scale of 1 to 5, ranging from worst to best. Detailed descriptions and settings for these metrics are provided in Appendix A.2.

4.1.4 Evaluation Methodology

In our study, we prompt 11 LLMs to generate answers to questions in the open-ended QA dataset. The objective was to generate five answers per question, with each answer strictly limited to no more than 150 words. This constraint was aimed to maintain focus and conciseness in the answers provided.

Following the answer generation phase, we utilized two advanced LLMs, GPT-4 and LLaMA3-70b, to evaluate the answers. These models were selected based on their proven capabilities in understanding and processing natural language, making them suitable for the task of assessing the quality of the answers generated by other LLMs.

However, there are some works (Bai et al., 2024) that raise significant concerns regarding the reliability of LLMs as evaluators. Their sensitivity to the specific textual instructions and inputs they receive can lead to inconsistencies. For instance, when the order of answers is altered during the evaluation process, it has been observed that the same model may provide different scores for the same set of answers. This variability indicates a potential vulnerability in the evaluation process, where the models could be manipulated to produce biased or unreliable evaluations.

To mitigate these challenges and enhance the reliability of our assessment, we have implemented

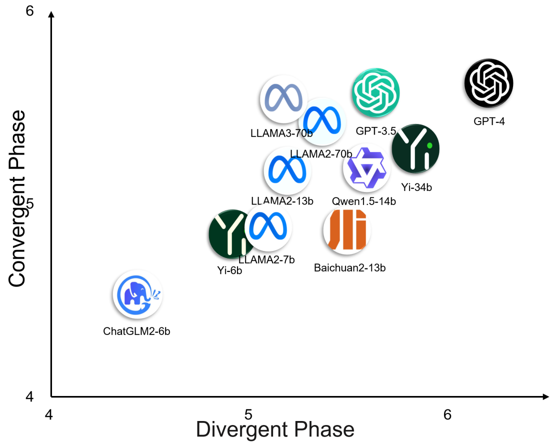

Figure 2: Visualization of LLM performance across divergent and convergent phases.

a refined approach involving pairwise comparison, specifically ranking, to enhance our assessment methodology. Instead of merely scoring the answers, each LLM (i.e., GPT-4, LLAMA3-70b) was also required to rank the answers within each group. This ranking process forces the models to directly compare answers against each other, which helps in reducing the impact of the order in which answers are presented. This method of pairwise comparison and ranking serves to standardize the evaluation process, ensuring that each answer is judged in relation to others in its group, thereby fostering a more consistent and fair assessment.

4.2 Main Experiment Results

In general, a comparison of the creative problem-solving abilities across different models reveals significant performance disparities. GPT-4 outperforms other models in both the divergent and convergent phases, underscoring its leading position in these tasks. Additionally, among models with similar parameter sizes—Yi-6b, ChatGLM-6b, LLAMA2-7b, and Qwen1.5-14b in one group, and BaiChuan2-13b and LLAMA-2-13b in another—there are notable performance variations within each group. This further validates the impact of model architecture on performance.

4.2.1 Model Performance Visualization

To thoroughly assess the correlation between divergent and convergent phases and overall model performance, we adopted the Analytic Hierarchy Process (AHP) as detailed in (Chulvi et al., 2013). This methodology allows us to compute weights and conduct consistency checks for the metrics associated with each phase. The specific computational steps are fully documented in Appendix A.3.

Model	Divergent Phase					Convergent Phase				
	Flu.	Nov.	Flex.	Rich.	Div. D.	P. S.	S. T.	D. M.	S. E.	Conv. D.
LLAMA-2-7b	7.88	6.82	7.26	7.08	3.21	6.83	6.41	6.56	6.42	3.61
LLAMA-2-13b	8.01	6.76	7.40	7.15	3.34	7.11	6.71	6.92	6.74	3.78
LLAMA-2-70b	8.12	6.85	7.77	7.40	3.46	7.41	6.85	7.30	6.88	4.07
LLAMA-3-70b	7.55	6.79	7.50	7.16	3.30	7.32	6.87	7.28	7.00	4.13
ChatGLM-6b	6.18	5.55	5.65	6.12	2.96	5.85	5.96	6.23	5.76	3.47
Qwen1.5-14b	8.11	7.41	**7.88**	7.18	3.71	7.15	6.99	7.08	7.00	3.89
Yi-6b	7.78	6.69	7.16	7.11	3.28	7.03	6.62	6.49	6.17	3.67
Yi-34b	8.06	7.41	7.50	7.84	3.93	7.26	6.90	7.04	6.93	3.91
BaiChuan2-13b	8.03	6.97	7.66	7.79	3.64	6.48	6.82	6.74	6.12	3.73
GPT 3.5	**8.23**	6.69	7.51	7.51	4.02	7.27	6.94	**7.39**	7.18	4.20
GPT-4	8.17	**7.54**	7.61	**8.07**	**4.76**	**7.58**	**7.08**	7.37	**7.27**	**4.41**

Table 2: Experiment results for LLMs in Open-ended question answering. Abbreviations used are: **Flu.** (*Fluency*), **Nov.** (*Novelty*), **Flex.** (*Flexibility*), **Rich.** (*Richness*), **Div. D.** (*Divergent Degree*), **P. S.** (*Problem Solving*), **S. T.** (*Strategic Thinking*), **D. M.** (*Decision Making*), **S. E.** (*Self Efficiency*), and **Conv. D.** (*Convergent Degree*). **Bold**: the best result; Underline: the runner-up result.

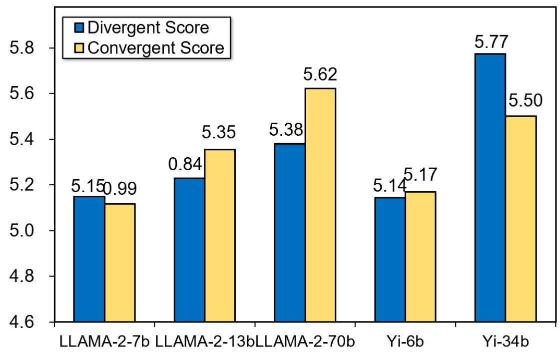

Figure 3: Relationship between models of the same series with different parameter sizes.

Utilizing these weights alongside macro indicators such as divergent degree and convergent degree, we have computed an aggregated evaluation index consisting of both a divergent score and a convergent score. These scores are visually represented as shown in Figure 2, which enables an intuitive comparison of different models' performances. The graph clearly demonstrates a positive correlation between the models' divergent and convergent capabilities, highlighting how strengths in one dimension often correspond to strengths in the other.

4.2.2 Impact of LLM Parameter Size

Analysis from the perspective of parameter size reveals a consistent trend, as demonstrated in the main experiment and detailed in Figure 3. Within the same architectural framework, there is a positive correlation between the performance of LLMs in both the divergent and convergent phases and their parameter size. This relationship suggests that as models increase in scale, their ability to handle complex creative problem-solving tasks improves significantly. This performance trend adheres to the scaling law (Kaplan et al., 2020), underscoring the critical role of parameter size in enhancing model capabilities. The correlation highlights the importance of scaling up models to achieve higher efficiency and effectiveness in creative tasks, thereby validating the scaling law's applicability to creative performance metrics in LLMs.

4.3 Domain-Specific Open QA Results

The comparative analysis across domains, as illustrated in Figure 4, underscores distinct domain-specific performances among the models. In domains like *General* and *Edu*, models such as GPT-4 consistently exhibit superior divergent and convergent phase scores, indicating a robust ability to generate novel ideas and connect disparate concepts effectively. Conversely, models like ChatGLM-6b show lower performance across most domains but notably lag in *Sci* and *Bio*, suggesting limitations in domains requiring highly specialized knowledge. The *Fin* domain presents a middle ground, with no single model dominating, reflecting a balanced

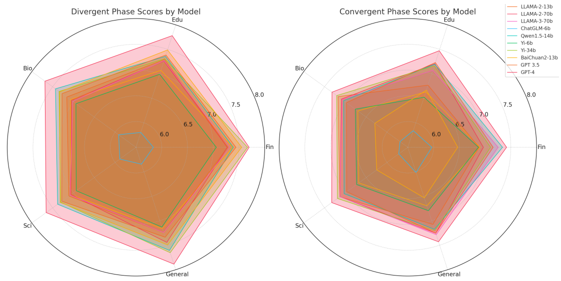

Figure 4: Radar charts displaying divergent and convergent phase scores of LLMs across five domains. Each plot illustrates domain-specific performance differences.

challenge in creativity and associative thinking tasks. These observations highlight the necessity of domain-specific tuning and evaluation to optimize models for varied Open QA applications.

4.4 DAT and RAT Experiments

This study extends its assessment of LLM creative capabilities by incorporating the DAT and the RAT, which evaluate divergent and convergent thinking abilities, respectively.

The DAT evaluates LLMs' ability to generate multiple creative ideas, focusing on fluency, flexibility, and originality. In this experiment, models are prompted to produce ten sets of unrelated nouns, totaling 100 groups. This format isolates semantic creativity by minimizing syntactic influence, ensuring the focus remains on the generative aspect of creativity. The DAT leverages datasets designed to elicit a high volume of diverse responses, consistent with benchmarks established in prior creativity research (Olson et al., 2021).

Conversely, the RAT assesses convergent thinking by challenging models to find connections among sets of three seemingly unrelated words and to generate a fourth word that links them all. This task tests the models' ability to synthesize and integrate disparate information into coherent outcomes. The RAT datasets are derived from classical studies (Bowden and Jung-Beeman, 2003), aligning the evaluation with well-validated measures of associative thinking.

Performance in the RAT is quantified by measuring the semantic distance between the model's output and the correct associative word from the dataset, providing a precise metric of associative accuracy. This measurement approach ensures a detailed and comparative analysis of the LLMs' proficiency in both generating novel ideas and synthesizing information.

For both the DAT and RAT tasks, metrics are directly adopted from previous studies (Olson et al., 2021; Mednick, 1962), using established benchmarks to maintain consistency with recognized methods in creativity assessment. The detailed formulation can be found in Appendix A.4

Model	DAT Score	RAT Score
LLAMA-2-7b	67.94	52.77
LLAMA-2-13b	70.37	46.81
LLAMA-2-70b	78.71	36.57
LLAMA-3-70b	77.85	35.15
Qwen1.5-14b	74.90	44.55
Yi-6b	67.01	55.39
Yi-34b	75.68	32.44
BaiChuan2-13b	71.64	46.33
Chatglm2-6b	62.13	58.06
GPT-3.5	82.10	30.95
GPT-4	**87.70**	**24.72**

Table 3: Results from DAT and RAT experiments.

The results from the DAT and RAT experiments reveal significant performance differences across models, highlighting their distinct capabilities in divergent and convergent thinking. GPT-4 excels in both tasks, reflecting its advanced ability to generate and connect ideas, likely due to its larger parameter size and advanced training. Conversely, LLAMA-2-70b and GPT-3.5 show a trade-off between high creativity and lower associative accuracy. Interestingly, the Chatglm2-6b scores suggest a specialization in associative thinking despite lower creativity scores. Overall, the performance trends observed here align with those seen in open-ended QA tasks, suggesting consistent model behaviors across different creative assessment.

5 Conclusion

This study presents a comprehensive benchmark that integrates an open-ended QA dataset with traditional creativity tasks, designed to assess the creative problem-solving abilities of LLMs across both divergent and convergent thinking phases. By employing multi-dimensional evaluation metrics, this benchmark effectively measures the capabilities of LLMs in relation to their architecture, parameter size, and domain-specific expertise, thereby advancing our understanding of creative cognition in AI and setting a new standard for evaluating AI creativity in fields.

6 Limitations

This study, while pioneering in its approach to evaluate creative problem-solving abilities of LLMs, acknowledges several limitations. Firstly, our exploration of creativity is confined to creative problem-solving within the scope of divergent and convergent thinking. Creativity is a multifaceted phenomenon that encompasses a broader spectrum of cognitive abilities and expressions which are not fully covered in this study. Further research is required to explore these dimensions comprehensively. Secondly, the current benchmarks, though effective, are primarily empirical. Future studies should aim to integrate theoretical frameworks or mechanisms that can provide deeper insights into the underlying processes that govern creativity in LLMs, thus enhancing our understanding and the evaluation of creative capacities in artificial intelligence.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, et al. 2021. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36.

Margaret A Boden. 1994. Creativity: a framework for research. *Behavioral and Brain Sciences*, 17(3):558–570.

Edward M Bowden and Mark Jung-Beeman. 2003. Normative data for 144 compound remote associate problems. *Behavior research methods, instruments, & computers*, 35:634–639.

Honghua Chen and Nai Ding. 2023. Probing the "creativity" of large language models: Can models produce divergent semantic association? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12881–12888.

Vicente Chulvi, María Carmen González-Cruz, Elena Mulet, and Jaime Aguilar-Zambrano. 2013. Influence of the type of idea-generation method on the creativity of solutions. *Research in Engineering Design*, 24:33–41.

J. Daniel Couger, Lexis F. Higgins, and Scott C. McIntyre. 1993. (un)structured creativity in information systems organizations. *MIS Q.*

David Cropley. 2023. Is artificial intelligence more creative than humans? : Chatgpt and the divergent association task. *Learning Letters*.

Zijian Ding, Arvind Srinivasan, Stephen MacNeil, and Joel Chan. 2023. Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. In *Proceedings of the 15th Conference on Creativity and Cognition*, pages 489–505.

Mark Du. 2023. Strategic thinking in artificial intelligence and expert: Problem solving and creativity.

Howard Gardner. 2011. *Creating minds: An anatomy of creativity seen through the lives of Freud, Einstein, Picasso, Stravinsky, Eliot, Graham, and Ghandi*. Civitas books.

Luis Fabricio Góes, Marco Volpe, Piotr Sawicki, Marek Grses, and Jacob Watson. 2023. Pushing gpt's creativity to its limits: Alternative uses and torrance tests.

Joy Peter Guilford. 2017. Creativity: A quarter century of progress. In *Perspectives in creativity*, pages 37–59. Routledge.

Erik E Guzik, Christian Byrge, and Christian Gilde. 2023. The originality of machines: Ai takes the torrance test. *Journal of Creativity*, 33(3):100065.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Joe Khatena and E Paul Torrance. 1973. Thinking creatively with sounds and words: Normstechnical manual. *Res. ed.) Bensenville, IL: Scholastic Testing Service*.

Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *ArXiv preprint*.

Sarnoff A. Mednick. 1962. The associative basis of the creative process. *Psychological review*.

Saeid Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. 2023. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. *Preprint*, arXiv:2306.11167.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van Maas. 2022. Putting gpt-3's creativity to the alternative uses test. *Preprint*, arXiv:2206.08932.

Douglas Summers-Stay, Clare R Voss, and Stephanie M Lukin. 2023. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. 2023. Macgyver: Are large language models creative problem solvers? *ArXiv preprint*.

E Paul Torrance. 1977. Creativity in the classroom; what research says to the teacher.

Donald J. Treffinger. 1998. Creativity, creative thinking, and critical thinking: In search of definitions. *Gifted and Talented International*.

Haonan Wang, James Zou, Michael Mozer, Anirudh Goyal, Alex Lamb, Linjun Zhang, Weijie J Su, Zhun Deng, Michael Qizhe Xie, et al. 2024. Can ai be as creative as humans? *Preprint*, arXiv:2401.01623.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, et al. 2024. Assessing and understanding creativity in large language models. *ArXiv preprint*.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2023. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. *arXiv preprint arXiv:2312.02439*.

A Appendix

A.1 Prompt Examples of Open-Ended QA Dataset Construction and Evaluation

A.1.1 Question Generation

Please generate a new situation in ***{Finance, Science, Education, Biology, General}*** *domain that has a very different* ***(stakeholder, context)*** *but very similar* ***(goal, obstacle)*** *based on the input:*

Input:

Stakeholder: a patient who has a malignant tumor in his stomach

Context: ray at low intensity is insufficient to destroy the tumor

Goal: destroy the tumor without affecting the healthy tissue

Obstacle: ray at high intensity will also destroy healthy tissue

A.1.2 Question Rewrite

You are a good writer. Please help me rewrite the given paragraph into a complete and coherent question. The rewritten question should include all the key points and details without introducing any additional information. Strive to make your rewritten content clear and concise. Paragraph: {original paragraph}

A.1.3 Answer Generation

You are an expert in ***{Finance, Science, Education, Biology, General}*** *domain, for a question, please give 5 creative solutions very concisely. Use as few steps as possible and each answer should ideally be less than 100 words. Question: {original question}*

A.1.4 Divergence Evaluation

You are a fair assessment expert, and you will be given one question along with 5 different answers. Your task involves evaluating answers using a set of specific criteria to ensure a fair and comprehensive assessment. Please follow these guidelines when scoring and ranking the answers:

a. Each answer should be evaluated in relation to its corresponding question. Assume your understanding of the question is correct for the purpose of this evaluation.

b. You should rate the answer on on four distinct metrics. Assign a score between 1 and 10, with 10 being the highest:

1. Fluency: Judge how smoothly and naturally the answer reads. Assess whether the language used is clear, engaging, and free from awkward phrasing or grammatical errors.

2. Novelty: Evaluate the originality of the content. Consider whether the answer provides unique insights or perspectives not commonly found in standard responses.

3. Flexibility: Determine the adaptability of the answer in addressing different aspects of the question. This involves considering whether the response can be interpreted positively in various contexts or under different assumptions.

4. Richness: Assess the depth and detail of the answer. Check whether it covers the subject comprehensively, including all relevant points and necessary explanations.

You should only give the score and the rank of each answer, Format like: Fluency: 3, Rank: 1. There is no need to explain the reasoning behind each score. After scoring and ranking, please provide a final score between 1 and 5 for the diversity of these five answers. Format like: diversity: 5

Important Note: Ensure that each score is based on the answer's own merits, not in comparison to other answers. The ranking should reflect the relative quality of the answers, but the scores should be fair and independent of each other.

Question: {Question} Answer1: {Answer1}Answer2: {Answer2} Answer3: {Answer3} Answer4: {Answer4} Answer5: {Answer5}

A.1.5 Convergence Evaluation

you are a fair assessment expert, and you will be given one question along with 5 different answers. Your task involves evaluating answers using a set of specific criteria to ensure a fair and comprehensive assessment. Please follow these guidelines to score and rank the answers:

a. Each answer should be evaluated in relation to its corresponding question. Assume your understanding of the question is correct for the purpose of this evaluation.

b. You should rate the answer on four distinct metrics. Assign a score between 1 and 10, with 10 being the highest:

1.Problem Solving: Assess how effectively the response addresses and resolves the core issue presented in the question. Consider the creativity and practicality of the proposed solutions.

2. Strategic Thinking: Evaluate the response's demonstration of long-term planning and foresight. Look for evidence of a thoughtful approach that considers various factors and potential outcomes.

3. Decision Making: Determine the decisiveness and rationale behind the choices made within the response. Assess how well the response justifies these decisions based on the information provided.

4. Self Efficiency: Judge the confidence and independence exhibited in the response. Consider how the responder demonstrates capability and resourcefulness in addressing the question.

You should only give the score and the rank of each answer, Format like: Problem Solving: 7, Rank: 1 There is no need to explain the reasoning behind each score. After scoring and ranking, please provide a final score between 1 and 5 for the convergence of these five answers. Format like: convergence: 4

Important Note: Ensure that each score is based on the answer's own merits, not in comparison to other answers. The ranking should reflect the relative quality of the answers, but the scores should be fair and independent of each other.

Question: {Question} Answer1: {Answer1}Answer2: {Answer2} Answer3: {Answer3} Answer4: {Answer4} Answer5: {Answer5}

A.2 Detailed Description of DAT and RAT task

A.2.1 Words Generation

Prompt of DAT: *Please write 10 nouns in English that are as irrelevant from each other as possible, in all meanings and uses of the words. Please note that the words you write should have only single word, only nouns (e.g., things, objects, concepts), and no proper nouns (e.g., no specific people or places).*

Prompt of RAT: *Please provide a word that is semantically related to each of the three terms I will give you, ensuring that the relationship is as close as possible to all three.*

A.3 Determining Weights of Indicators using Analytic Hierarchy Process (AHP)

To determine the relative importance of various indicators in both the Divergent Phase and Convergent Phase of our study on creative problem-solving, we employed the Analytic Hierarchy Process (AHP). Below, we detail the steps taken to derive the weights for each indicator, ensuring consistency in our judgments.

A.3.1 Divergent Phase

The indicators for the Divergent Phase were: Fluency, Novelty, Flexibility, and Richness. We conducted pairwise comparisons of these indicators to construct the judgment matrix, followed by consistency analysis and adjustment.

Pairwise Comparisons

$$\mathbf{A}_{\text{divergent}} = \begin{pmatrix} 1 & \frac{1}{3} & \frac{1}{2} & \frac{1}{2} \\ 3 & 1 & 3 & 2 \\ 2 & \frac{1}{3} & 1 & \frac{1}{2} \\ 2 & \frac{1}{2} & 2 & 1 \end{pmatrix} \tag{1}$$

Priority Vector and Consistency Ratio Using the principal eigenvector method, we obtained the priority vector and checked the consistency ratio (CR).

$$\begin{aligned} \text{Priority Vector:} \quad & [0.1190, 0.4512, 0.1689, 0.2609] \\ \text{Max Eigenvalue:} \quad & \lambda_{\max} = 4.0710 \\ \text{Consistency Ratio (CR):} \quad & CR = 0.0260 \end{aligned}$$

Since the CR is less than 0.1, the consistency of our judgment matrix is acceptable.

A.3.2 Convergent Phase

The indicators for the Convergent Phase were: Problem Solving, Strategic Thinking, Decision Making, and Self Efficiency. Similar steps were followed as in the Divergent Phase.

Pairwise Comparisons

$$\mathbf{A}_{\text{convergent}} = \begin{pmatrix} 1 & 3 & 3 & 2 \\ \frac{1}{3} & 1 & 1 & \frac{1}{2} \\ \frac{1}{3} & 1 & 1 & \frac{1}{2} \\ \frac{1}{2} & 2 & 2 & 1 \end{pmatrix} \tag{2}$$

Priority Vector and Consistency Ratio

$$\begin{aligned} \text{Priority Vector:} \quad & [0.4554, 0.1409, 0.1409, 0.2628] \\ \text{Max Eigenvalue:} \quad & \lambda_{\max} = 4.0104 \\ \text{Consistency Ratio (CR):} \quad & CR = 0.0038 \end{aligned}$$

After adjustments, the CR is less than 0.1, indicating acceptable consistency in our judgments.

A.3.3 Conclusion

The AHP method allowed us to systematically derive the weights for the indicators in both the Divergent and Convergent Phases, ensuring that our judgments were consistent and reliable. The final weights for each phase are as follows:

- **Divergent Phase:**
  - Fluency: 0.1190
  - Novelty: 0.4512
  - Flexibility: 0.1689
  - Richness: 0.2609
- **Convergent Phase:**
  - Problem Solving: 0.4554
  - Strategic Thinking: 0.1409
  - Decision Making: 0.1409
  - Self Efficiency: 0.2628

These weights were then used to evaluate and compare the creative problem-solving capabilities of the models under study.

A.4 Formulation of DAT and RAT

The DAT score, for instance, is calculated as the average cosine distance between word embeddings of the given nouns, formalizing the evaluation of the models' creative output.

$$\text{DAT} = \frac{100}{n(n-1)} \sum_{\substack{i,j \\ i \neq j}}^{n} (1 - \cos(w_i, w_j)) \tag{3}$$

Similarly, given n samples, denote the generated word embeddings w_i and the label word embeddings l_i, the RAT score can be calculated as follows:

$$\text{RAT} = \frac{100}{n} \sum_{i}^{n} (1 - \cos(w_i, l_i)) \tag{4}$$